ONE DOCUMENT AT A TIME: SMALL SCALE DIGITZATION PROJECTS

Peter Brueggeman
Scripps Institution of Oceanography Library
University of California, San Diego
La Jolla California 92037 USA

Janet Webster
Guin Library, Hatfield Marine Science Center
Oregon State University
Newport, Oregon 97365 USA

Barbara Butler
Oregon Institute of Marine Biology
University of Oregon State University
Charleston, Oregon 97420 USA

Abstract: The members of this panel have been involved in small-scale digitization and each has taken a different approach. Though we vary in strategy and processes, we have found that digitization and archiving can be accomplished even on a very tight budget, and can be juggled into your workday if need be. Our experiences demonstrate that other IAMSLIC members can dive into their digitization interests right from their desks.

Keywords: digitization standards; scanning technology; collection development; Aquatic Commons.

Part 1: Legacy Publication Digitization at Scripps

Scripps Institution of Oceanography has a long history of publishing research monographs and technical reports in print, now almost entirely discontinued. Scripps Library is digitizing this back stock of legacy print publications either on demand by fulfilling requests, or by digitizing items of enduring value. Past text digitization of these legacy publications by Scripps Library have largely involved vendor production of PDFs from scanning, optical character recognition, and then XML-TEI encoded text production, followed by PDF creation. This process delivers PDFs of the smallest file size, analogous to contemporary PDF production by electronic journal publishers, but is relatively costly to produce and involve considerable time in vendor interaction, proofing, and revisions. Scripps Library also experimented with using its interlibrary loan (ILL) staffing to produce PDFs as part of their work stream. This proved problematic. It was

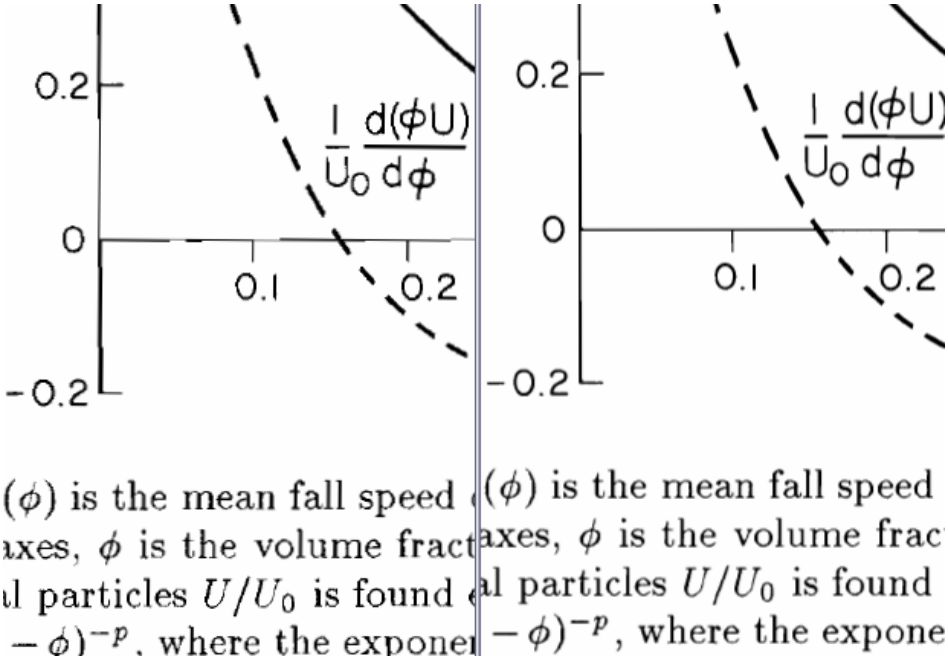
difficult to insert additional work of a low priority into a busy workflow. ILL PDF production uses a scanning resolution suitable for ILL but not as high a resolution as deemed suitable for producing online versions of legacy publications. There were also proofing issues in that ILL staff go quickly through their work and scanning errors can get by without detection. So a do-it-yourself desktop approach was settled upon to achieve high quality results, at reduced cost, and least effort.

The current implementation uses Adobe Acrobat Professional to create PDFs from a Hewlett Packard ScanJet 7800 sheet-feeding scanner. The HP ScanJet 7800 scanner can scan up to fifty pages on both sides (duplex scanning); it will scan more than fifty pages at a time, but feeding problems may arise. A Plustek OpticBook 3600 Corporate flatbed book scanner is used to scan publications with tight bindings that cannot be unbound. The OpticBook scanner can scan up to six millimeters from the flatbed edge, which is perfect for publications with tight bindings and narrow gutters that cannot be disbound. After scanning, optical character recognition is run on the scanned-page PDF using Adobe Acrobat, so that searchable text is hidden behind each scanned page image in the PDF.

Scanning is accomplished from disbound, trimmed original publications in order to take advantage of automated scanner sheetfeeding. Scripps Library maintains a back stock of legacy publications for such digitization. The glued or stitched binding is cut off with a clamping paper cutter, or if stapled, the staples are removed and the fold between successive pages trimmed off. After trimming the document into individual pages, the entire document is checked to ensure each page is separate so that there are no misfeeds in the scanner. If an original publication is not available to disbind for scanning, then the text-only pages are photocopied and run through the sheet-feeding scanner.

A few pages are scanned first as a test, to see if results are up to expectations. For original documents that are yellowed or browned, the lightening setting in the scanning software is adjusted to produce whiter pages.

Scanning of text-only pages including those with black/white line drawings, graphs, and figures is done at 600 pixels per inch (ppi) black/white text scanning (two bit). Scanning at 600ppi gives very sharp-looking text, which can be assessed by zooming such a PDF up to 200%. However photographs and half-tone images are not scanned at this resolution due to PDF file size concerns. Scanning text-only pages at 600 ppi is much higher resolution than seen in some scanned PDF projects. The intention is to produce a very high quality legacy PDF publication, and not do it again in the future as opinions about resolution and quality and PDF file size may change (as they have done in the past). A 600 ppi produced PDF continues to be sufficiently small in file size even with such high resolution text scanning, and the slower scan time at this relatively high resolution is not a problem when using a sheet feeding scanner in a non-production desktop environment.



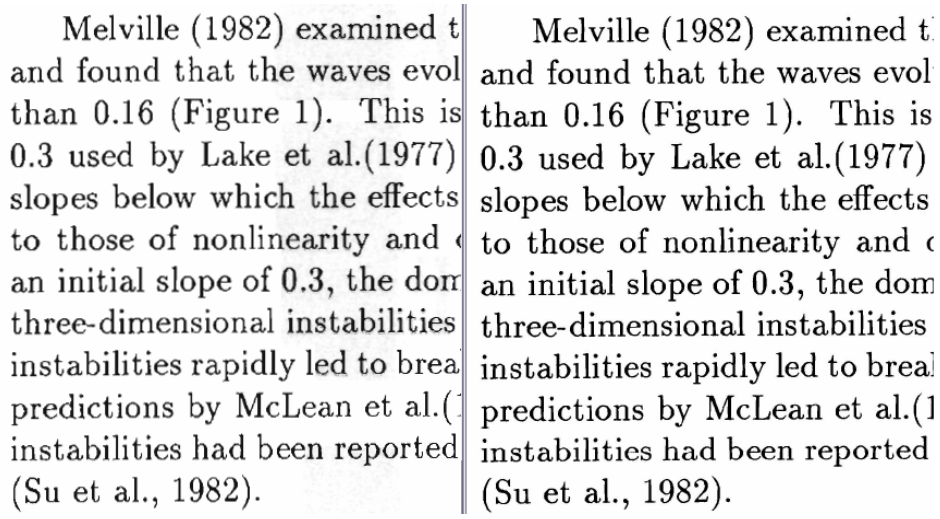
The above figure shows 300 ppi black/white text scanning on the left and 600 ppi black/white text scanning on the right, for a PDF zoomed up to 200%. Note the sharper axes of the graph for the 600 ppi scan on the right, as well as the sharper appearance of its text characters.

1. Four pages of text scanned at 300 ppi black/white results in a PDF of 175K in file size.
2. Four pages of text scanned at 600 ppi black/white results in a PDF of 328K in file size, a bit less than double.
3. A forty-page text document would be 1.75 megabytes if scanned at 300 ppi black/white.
4. A forty-page text document would be 3.28 megabytes if scanned at 600 ppi black/white.

This increased file size at 600 ppi black/white compared to 300 ppi black/white is relatively modest for the benefit of sharper text and figures throughout the entire document.

300 ppi grayscale or color scanning is used for pages containing black and white or color halftone photographs respectively. 300 ppi is used for grayscale or color scanning instead of 600 ppi because the resulting PDF just gets too large in file size if photographs are scanned at 600 ppi. Grayscale or color scanned pages consume considerable file size in a final PDF, and so grayscale or color scanning is not used for text-only pages.

1. Four pages of text scanned at 600 ppi black/white results in a PDF of 328K in file size.
2. Four pages of text scanned at 300 ppi grayscale results in a PDF of 1,150K in file size, which is 3.5 times more file space for half the resolution.

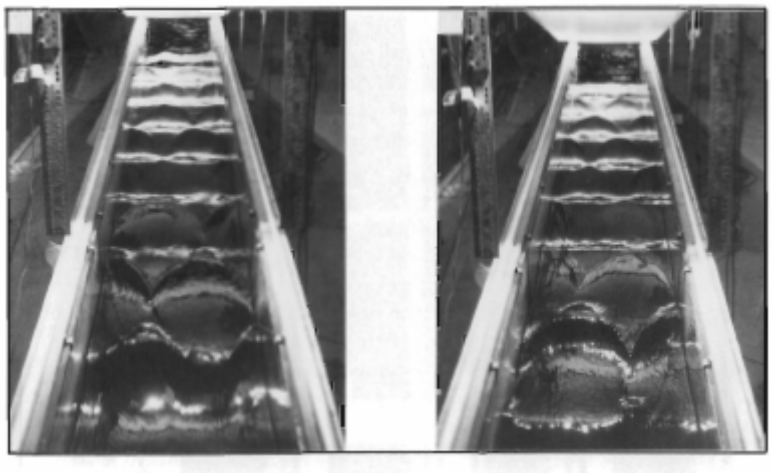


Melville (1982) examined t
and found that the waves evol
than 0.16 (Figure 1). This is
0.3 used by Lake et al.(1977)
slopes below which the effects
to those of nonlinearity and
an initial slope of 0.3, the dom
three-dimensional instabilities
instabilities rapidly led to brea
predictions by McLean et al.(
instabilities had been reported
(Su et al., 1982).

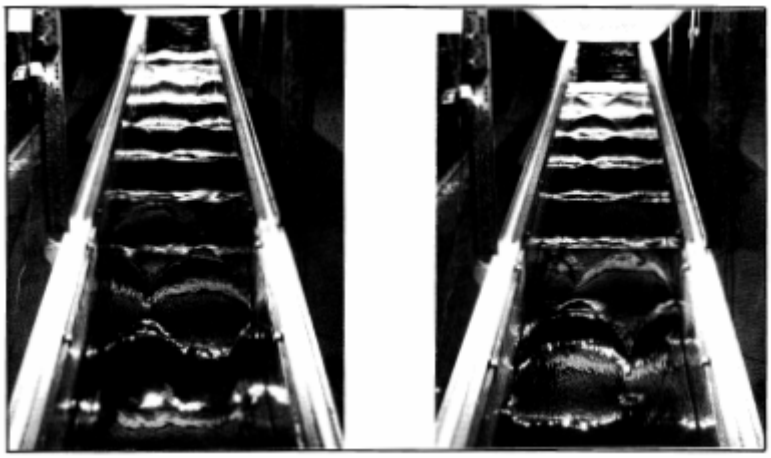
Melville (1982) examined t
and found that the waves evol
than 0.16 (Figure 1). This is
0.3 used by Lake et al.(1977)
slopes below which the effects
to those of nonlinearity and
an initial slope of 0.3, the dom
three-dimensional instabilities
instabilities rapidly led to brea
predictions by McLean et al.(1
instabilities had been reported
(Su et al., 1982).

In addition to file size differences between 300 ppi grayscale scanning and 600 ppi black/white scanning of text-only pages, the above figure compares 300 ppi grayscale scanning on the left with 600 ppi black/white text scanning on the right, for a PDF zoomed up to 200%. It can be seen that grayscale scanning doesn't produce clear and crisp text compared to black-white scanning. There are gray shadings on what should be a clean white background (above and below the word "nonlinearity"), and 300 ppi grayscale text is not quite as sharp at 600 ppi black/white text (compare the words "three", "by", etc.).

If a page is composed of a halftone photograph with a large amount of text, some file space can be saved by pasting a cropped grayscale or color scan of the photograph onto the black/white scanned text page. This combines the sharp appearance of text scanned at 600 ppi, with the higher quality of halftone photos scanned in grayscale or color.



Black and white photographs scanned at 300 ppi grayscale.



Black and white photographs scanned at 600 ppi black/white have lost detail compared to the 300 ppi grayscale scan above.

First the page is scanned at 600 ppi black/white and saved as a TIF. Then the page is scanned 300 ppi grayscale or color but framed for the halftone photograph. Then the scan is cropped closely around the photograph, which is then pasted over the lower quality photograph on the black/white scanned TIF. Then this composite TIF is inserted into the PDF. This works best where the page is mostly text with the black and white or color photograph being a small element on the page.

1. A single page with a black and white halftone photograph scanned at 600 ppi black/white gives a PDF of 170K size, with an unacceptable appearing photograph as explained above.

2. A single page with a black and white halftone photograph scanned at 300 ppi grayscale gives a PDF of 760K size and a good looking photograph but the text is less than acceptable in appearance as explained above.
3. A single page with a black and white halftone photograph and scanned at 600 ppi black/white for the text and 300 ppi grayscale for the pasted-in photograph gives a PDF of 1,275K size. The text is sharp in appearance and the photograph shows detail well.
4. A single page with a black and white halftone photograph and scanned at 600 ppi grayscale gives a PDF of 1,436K size. Text and photograph look great, but considerable file size is consumed if the original has many pages with photographs.

So scanning and creating such a composite page combining black/white scanned text and a grayscale or color scanned photograph consumes an additional 515K in file size per page compared to a 300 ppi grayscale/color scanned page. The composite page has better appearing text and a good quality photograph in appearance. This is a modest increase in file size IF such pages are not numerous. However an additional 5 megabytes is added for every ten such composite pages. Therefore for documents with a lot of pages with text and photographs, it is best to scan those pages at 300 ppi grayscale or color, and scan the text-only pages at 600 ppi black/white, accepting the slightly degraded appearance of the text on those grayscale- or color-scanned pages with photographs.

After scanning the original document in groups of fifty pages, the various PDFs are assembled into one PDF. Then Adobe Acrobat's optical character recognition software is run against the PDF to create a word or phrase searchable PDF (via Adobe Acrobat's DOCUMENT - RECOGNIZE TEXT USING OCR.) Adobe Acrobat's OCR software is not highly accurate, and its percentage of success depends on the typescript quality in the original. As a result, if searching a word(s) or a phrase within a PDF yields zero results for that PDF, it cannot be assumed that such a word(s) or phrase does not exist within that PDF. However Adobe Acrobat's OCR is sufficiently successful for document discovery searching via Google, since key words are typically used many times within a document. Other optical character recognition could be investigated if improved OCR was important.

Considerable file space can be wasted during assembly of scanned pages into the final PDF. This is revealed through Adobe Acrobat's ability to analyze file space consumed by components of a PDF file (via ADVANCED – PDF- OPTIMIZER - AUDIT SPACE USAGE). "Document Overhead" can sometimes, but not always, be a considerable percentage of the final PDF file size, and is related to a number of things that make up the structure of a PDF file. Document Overhead can reduced significantly in some but not all cases, so it is useful to attempt to reduce it. First save the final PDF, and then save it again through FILE - SAVE AS, using the same file name if you wish. If PDF document overhead is being inordinately consumed, this FILE – SAVE AS procedure will reduce it, and thus the PDF file size will be reduced.

Page proof the final PDF by paging through it quickly to ensure all pages are present. Count off the page numbering as you click-page through the PDF, since sheet feeding scanners can pass two pages through at a time if they are stuck together.

If the final PDF is too large in file size, compress the PDF. Compression will alter the quality of viewed images and page scans, so it should be used judiciously. It is important to keep the original uncompressed PDF, in case it is necessary to go back to it and start over with a different compression method. At Scripps Library, PDF file sizes under twenty megabytes seem to deliver best with the underlying software for the University of California's eScholarship Repository and we don't have bandwidth limitations, so achieving a small PDF file size is not a great concern for us. Your mileage may vary. To compress a PDF, there is advice on the Web, which you can find via Google searching.

You can run Adobe Acrobat's ADVANCED - PDF OPTIMIZER, and try its default compression setting for IMAGES. Look at the resulting PDF and file size, and it may meet your needs. You can try compression for SCANNED PAGES; select "Optimize Compression of Page Regions Based On Color Content' and slide the slider knob between "small size" and "high quality." Start out by choosing something between half and three-quarters on the slider, depending on how much compression being sought for a targeted file size. Look at your resulting PDF and its file size. Produce successive PDFs trying a few different positions on this slider, and do not over-compress. Remember to always keep an uncompressed version of your PDF; don't overwrite it. You may wish to serve that uncompressed PDF in the future as bandwidth becomes less of an issue. You may need to execute a less aggressive compression later if you don't recognize a problem immediately with your current PDF compression method.

Get going on those legacy publications from your institution, one at a time, no rush... Time will pass and much will get done.

Part 2: A Cog in the OSU Libraries Digitization Process

The Oregon State University Libraries are actively engaged in building digital collections. Starting with a rich resource of the Linus Pauling Papers in our Special Collections, we have expanded to digitizing other parts of the collection as well as identifying material beyond our own holdings. It takes many cogs in this digitization machine to make the process work smoothly and produce a useful product. The following describes the role of the subject librarian as an important cog whose primary role is in identifying what to digitize. This is more about selection than scanning.

The Digitization Process:

The librarian is one cog in the OSU digitization machine. She identifies possible candidates for scanning, investigates and clears copyright, selects the appropriate digital

collection, decides on the disposition of the material, and then passes it on to the Digital Production Unit (DPU). Once there, the material is prepared for scanning. This may mean disbinding if appropriate. The librarian has already decided what should happen to the material once scanned: rebind and re-shelve, box and re-shelve, or discard. When possible, we look for duplicates to scan so we can maintain anything already catalogued in the collection and then discard the duplicate after digitization. Once prepared, the material is scanned and OCRed. The DPU staff members check for quality and then create the metadata record for the material. They deposit it into the selected collection. The DPU staff includes library technicians and students workers with supervision by a librarian.

The DPU has an excellent wiki site that contains all needed information and documentation for the librarian and the staff (http://wiki.library.oregonstate.edu/confluence/display/TechServ/Digital+Production+Unit+Documentation). The librarian finds necessary forms to include with all candidates for digitations. The staff locates useful data dictionaries. General information is maintained on scanning standards. OSU uses DSpace and ContentDM to manage most of its digital assets. Our DSpace instantiation has significant revisions and additions. The most developed is the metadata schema for the OregonExplorer as this contains very useful pull down menus for spatial data from broad descriptors such as county or basin to hydrologic unit codes.

How the Process Works:
Three examples illustrate how the OSU process works from the librarian's perspective.

Some collection development grows out of ongoing projects. I have been working with a citizens group concerned with one of Oregon's smaller estuaries, Netarts Bay. They wanted better access to reports and documents so contacted me for assistance recognizing that OSU probably had much of the desired material. I started by compiling a bibliography of research and historical documents on the bay. I shared this with the group who identified the high priority digitization candidates. One document was an easy choice: *The natural resources and human utilization of Netarts Bay, Oregon* edited by Heather Stout, 1976. It was an NSF funded student project, we had multiple copies in the library and I happened to know Ms Stout as she works at my institution. She had an extra copy that she was willing to sacrifice for the greater good and she could apprise me of the copyright standing. This is an example of easy it can be to select material, find a duplicate, clear the copyright and send off to processing.

This second example is more complex. This dissertation was identified through the Netarts project again: *Tillamook prehistory and its relation to the Northwest coast culture area* by Thomas M. Newman, 1959. This was another high priority as it is one of the few books to detail early archaeology of the Oregon coast. While selecting it was logical, the problems with scanning it were multiple.

- It was not on OSU publication, so I needed to validate why we should add it to OSU's repository (regional interest and unique coverage).
- It was not in the public domain, so copyright permission needed to be secured.
- Finally, it originally appeared as a University of Oregon dissertation, but was republished as a monograph by the University of Oregon's Department of Anthropology. Consequently, I had to determine if there were major differences between the two documents.

As the item originated from the University of Oregon, I asked Barbara Butler if she could track down the original dissertation and compare it to the monograph. She worked with the Anthropology Department who were very helpful, and we decided that the monograph was acceptable to scan. Barbara and I worked together to track down the author's widow who gave us permission to scan and post. I requested that the OSU copy be scanned and then boxed for shelving. This example illustrates the power of collaboration and shared problem solving.

The final example is a work in progress. Recently, a retired OSU faculty member donated a long run of the local journal, *Oregon Birds*. I posted the gift to the Cyamus discussion list to see if anyone needed copies, and several people suggested that it be digitized. Consequently, I contacted the publisher, the Oregon Field Ornithologists who were interested and have current issues in PDF format. I generated a project budget with help from my Technical Services Department chair and proposed an $8,000 project to the OFO. The cost includes staff time to scan, OCR and catalog at the issue level as each issue has unique content on different species. An added twist is how to handle the ongoing publication of the journal and its digital archiving. My roles throughout are to identify the value to the OSU Libraries and its user community, and then negotiate with OFO.

Why this Process Works:

Everyone can contribute to the digitization of the grey literature of marine and aquatic science. In my case, I concentrate on identifying and selecting material to build coherent collections as well as provide electronic access to important regional items. I know my limitations and that I cannot do the whole process at my branch library. I have access to a very good digitization unit. We have an established workflow and means to get material into that workflow. I use the workflow; I promote it and thank those involved. I work

with others both internally and externally to utilize this expertise. By contributing at my local level, I contribute to the greater community. We all can do that.

Part 3: Oregon Institute of Marine Biology (OIMB) and digital repositories

My situation at OIMB is a bit different than Scripps and OSU. I created an OIMB Community within University of Oregon's Dspace based digital repository, Scholars' Bank (https://scholarsbank.uoregon.edu/dspace/handle/1794/516). Like Janet, I am responsible for building a digital collection, selecting materials and assuring that copyright clearances or permissions have been obtained. However, University of Oregon does not have staff devoted to this project, so I digitize documents on my own. The "collections" within the "community" include: coastal gray literature, Coos Watershed Association, OIMB class photos, OIMB publications, student reports, theses, and South Slough National Estuarine Research Reserve.

My initial focus has been to archive locally produced documents such as student reports, theses and dissertations. I try to obtain these materials in electronic format from the students, but scan those not available to me digitally. I initially produced PDFs using my interlibrary loan flatbed scanner and Ariel software. Colleagues on main campus performed OCR and then posted the documents to the repository on my behalf. I tried scanning documents at 600ppi as suggested, but found the process quite slow and opted to scan at 300ppi rather than invest additional time. I have now acquired an Epson GT-2500 document scanner and can make use of either the native scanner interface or Ariel scanning software. I do not have the option of dithering scans, so for pages with images I toggle back and forth to grayscale scanning (still 300ppi), or insert separately scanned pages after the fact using Adobe Acrobat Professional. I now perform OCR and deposit items to the repository myself. A student occasionally assisted me and could scan roughly 100 pages per hour. As the only staff member I checked all work for accuracy.

An example of the type of locally produced documents I am archiving is the out-of-print *Oregon Estuarine Invertebrates.* This was an ideal candidate for the OIMB Collection. I lobbied to have a link to the Scholars' Bank archive from our catalog record (not standard UO practice). Before this link was created you could find the digital version I created by searching Google, but not by searching our catalog. By collaborating with Janet, whose university takes a different approach to metadata, linking and digitization processes, I can effectively argue for different practices within my own institution. As Janet and I continue to collaboratively digitize coastal gray literature I may eventually begin to archive materials in an OSU based repository such as OregonExplorer if the material falls outside of the scope of the OIMB Community in Scholars' Bank.

The importance of collaborative work can't be ignored. Another item I have digitized is *The History of the Port of Coos Bay, 1852-1952*, a Pan American University thesis by George Case from 1983. Mr. Case granted me permission to include his thesis within our repository and even provided me with an unbound copy to digitize. The same day I digitized this document Mr. Case was asked by Southern Oregon University if they might include his thesis in their repository. We need to communicate our efforts widely or risk duplicating efforts.

A final example of my digitization efforts is *Laboratory and Field Text in Invertebrate Zoology*. This is the 1941 precursor to *Light's Manual*, is within the public domain and is only held by fifteen libraries worldwide (according to OCLC) and three IAMSLIC libraries (according to the IAMSLIC Z39.50 Distributed Library). However, it falls outside the scope of the OIMB collection so has been scanned and deposited within the Aquatic Commons.

My primary focus continues to be locally produced materials appropriate to the OIMB Community within the UO repository, but I will continue to alert Janet if materials are more appropriate for the OSU repository and will continue to contribute to Aquatic Commons.

If I can do this in my one-person library you can too.